

Analysis of Partially Observed Networks via Exponential-family Random Network Models

Ian E. Fellows

Mark S. Handcock

Department of Statistics, University of California, Los Angeles, CA 90095-1554

Summary. Exponential-family random network (ERN) models specify a joint representation of both the dyads of a network and nodal characteristics. This class of models allow the nodal characteristics to be modelled as stochastic processes, expanding the range and realism of exponential-family approaches to network modelling. In this paper we develop a theory of inference for ERN models when only part of the network is observed, as well as specific methodology for missing data, including non-ignorable mechanisms for network-based sampling designs and for latent class models. In particular, we consider data collected via contact tracing, of considerable importance to infectious disease epidemiology and public health.

1. Introduction

It is not uncommon for researchers to collect data on a subset of a single network rather than observing the full network. This partially observed case has been studied within the framework of exponential-family random graph models (ERGM) by Handcock and Gile (2010), however their formulation suffers from the limitation that any nodal attributes included in the model must be fully observed, and only dyads may be missing. This assumption is not met in most sampling designs, where only some of the nodes are surveyed by the researcher, and reduces the practical usage of ERGMs in the missing data setting.

By including nodal attributes as variates rather than fixed quantities, exponential-family random network models (ERNM, Fellows and Handcock, 2012) can provide a convenient basis for inference in cases where the data is partially unobserved, either due to design, or out-of-design (e.g., non-response) mechanisms. While our framework is applicable to all partial observation mechanisms we consider three common mechanisms for partial observations in more detail, specifically:

Missing Data: If the population is comprised of a large number of units, or the number of edges is large, it is relatively common to find that the resources to observe a full network are not available. Often units or dyads are unavailable for sampling or do not provide complete responses to a survey instrument. In this case, only some of the dyads and nodal characteristics are collected. We treat missing data as a form of sampling in which the sampling mechanism is unknown and outside the control of the researcher, or an *out-of-design missing data mechanism*. A good example of this is the National Longitudinal Study of Adolescent Health (Add Health), a school-based, longitudinal study of the health-related behaviours of adolescents and their outcomes in young adulthood. The study design sampled 80 high schools and 52 middle schools from the U.S., representative with respect to region of country, urbanicity, school size, school type, and ethnicity (Harris et al., 2003). In 1994-95 an in-school questionnaire

was administered to a nationally representative sample of students in grades 7 through 12. In addition to demographic and contextual information, each respondent was asked to nominate up to five boys and five girls within the school whom they regarded as their best friends. Thus each student could nominate up to ten students within the school (Udry, 2003). The nominations and contextual information were not available for some of the adolescents, either due to absence from school while the survey was being conducted, or refusal to participate. Thus, both the graph and nodal variates contained missing values.

Network sampling designs: Many studies in hard to reach populations use study designs that trace the linkages of an underlying social network. In these designs, the network is partially observed, however it is not of primary interest to the researcher. Such sampling designs have been exploited to estimate population disease rates (Gile and Handcock, 2010; Gile, 2011; Gile and Handcock, 2011).

Latent variables: Some quantities of the network may be in principle unobservable. The probability model for a network may posit the existence of unknown variables which do not correspond to any observable quantity. For example, stochastic block models (Nowicki and Snijders, 2001) posit the existence of classes of nodes, conditional upon which the dyads are independent. These classes are unobservable nodal characteristics and must be inferred from the relational data. Similarly, latent position cluster models (Handcock et al., 2006) posit the existence of unobservable continuous nodal quantities that provide a spatial geometry for the network structure.

In this paper we develop approaches for each of these scenarios in the context of ERNMs. Sections 2 through 4 introduce ERNM and extend the theory to incorporate partially observed populations. Section 5 develops methodology for each of the scenarios. Sub-section 5.1 looks at the effect of random non-response, and sub-section 5.2 applies a latent class model to extract unknown clusters from a real data-set. Sub-section 5.3 develops estimates based on contact tracing designs, which is of vital importance to the public health community. To our knowledge, the methods outlined in this paper represent the first statistically justifiable approach to inference from contract tracing data.

2. Exponential-family random network models

Exponential-family random network models (Fellows and Handcock, 2012) are a generalisation of the exponential-family random graph model (Frank and Strauss, 1986; Hunter and Handcock, 2006), where both dyads and nodal characteristics are treated as random variates. Formally, in a population of n units, let $Y_{i,j}$ indicate that unit i has a tie to unit j . Let Y be an $n \times n$ matrix $[Y_{i,j}]$ and X be an $n \times K$ matrix $[X_{ik}]$ of unit covariates. We define a network T as the union of the nodal covariates and the graph structure (i.e. $T = \{X, Y\}$). An exponential family model of T is expressed as

$$P(T = t|\eta) = \frac{1}{c(\eta, \mathcal{T})} e^{\eta \cdot g(t)} \quad t \in \mathcal{T}, \quad (1)$$

where $\eta \in R^q$ is a vector of parameters, g is a q -vector valued function defining a set of sufficient statistics, \mathcal{T} is the sample space of networks and $c(\eta, \mathcal{T}) = \sum_{t \in \mathcal{T}} e^{\eta \cdot g(t)}$ is the normalising constant. This model is developed in Fellows and Handcock (2012).

2.1. The Simple Homophily Model

Though any set of network statistics can be represented by g in equation (1), the examples in this paper will focus on a particularly parsimonious, but powerful, network model. Suppose that $X = (X_1, \dots, X_n)$ is a univariate categorical variable with m levels, labelled $0, \dots, m-1$. If $X_i = l$ we say that unit i is in group l . A joint model for X and Y is

$$P(T = (y, x) | \eta) = \frac{1}{c(\eta, \mathcal{T})} e^{\eta_0 \sum_{i,j} y_{i,j} + \eta_2 h(y, x) + \sum_{l=0}^{m-2} \eta_{j+3} \sum_{i=1}^n I(x_i = l)} \quad (y, x) \in \mathcal{T}.$$

The first term of this model is the number of edges, and controls the density of the graph. The last term represents the number of nodes in each category of x , except for the last level, which is dropped to maintain identifiability of the model. The second term h is the regularised sample homophily of x , as introduced by Fellows and Handcock (2012), and is defined as

$$h(y, x) = \sum_{k=0}^{m-1} \sum_{i: x_i = k} \sqrt{d_{i,k}(y, x) - E_{\perp}(\sqrt{d_{i,k}(y, x)})},$$

where $d_{i,k}(y, x)$ is the number of edges between node i and nodes in group k , and $E_{\perp}(f(Y, X))$ is the expectation of the statistic $f(Y, X)$, conditional upon $Y = y$ and the category counts (that is, the number of nodes in each category of x , $n(x) = \{n_k(x)\}_{k=1}^K$), assuming that X and Y are independent. Thus, each term in the sum is the square root of the number of neighbours of a node which share the same category, minus what would be expected by chance. Using this form of homophily avoids the degeneracy problems found in other formulations. For a more thorough justification, see Fellows and Handcock (2012).

While the examples in this paper focus on applications of the simple homophily model, the framework presented here applies to any arbitrary set of network statistics g . For example, in many applications the nodal attributes are multivariate, and their relationships are of interest to the researcher. Fellows and Handcock (2012) developed a network statistic that can be interpreted as a conditional logistic regression term which, if included, can model the relationship of several categorical variates.

3. Likelihood-based Inference from Partially Observed Networks

In this section we develop likelihood-based inference for network models based on partial observation of the networks. The approach allows non-ignorable sampling mechanisms for the networks, including some common network-based sampling designs.

Handcock and Gile (2010) developed a theory of missing data for ERG models, and the specification for ERN models proceeds similarly, though our formulation supports a more general class of missingness processes known as missing not at random (MNAR; see Rubin, 1976). Let T_{obs} and T_{miss} represent, respectively, the observed and unobserved part of the complete network T . We write $T = (T_{obs}, T_{miss})$, with realisations $t = (t_{obs}, t_{miss})$. Let W be a random variable representing the sampling process with realisation w . The probabilistic distribution of W is the *sampling mechanism*, and must fully specify the sample selection process, including the partition of T into T_{obs} and T_{miss} . Typically, W will consist of an n by n matrix indicating whether the dyad was sampled, and an n by K matrix indicating which nodal attributes are missing; however, W may contain additional information about the sampling, such as the order of sampling.

We write the *full data likelihood* as

$$p(T = t, W = w|\eta, \theta) = p(W = w|T = t, \theta) \frac{1}{c(\eta, \mathcal{T})} e^{\eta \cdot g(t)},$$

and we wish to draw inferences about η from the *observed data likelihood*, defined as

$$p(T_{obs} = t_{obs}, W = w|\eta, \theta) = \sum_{t_{miss}} p(W = w|T = (t_{obs}, t_{miss}), \theta) \frac{1}{c(\eta, \mathcal{T})} e^{\eta \cdot g((t_{obs}, t_{miss}))}. \quad (2)$$

This probability model jointly represents the distribution of the network T , and the sampling process W . The functional form of $p(W = w|T = t, \theta)$ is dependent on the form of missingness, and will differ depending on how T_{obs} was obtained. Section 5.3 illustrates a design of particular interest known as biased seed link tracing. When the sampling probabilities only depend on the observed data, then the sampling design is *amenable* to the model (Handcock and Gile, 2010), and is ignorable in the sense of Rubin (1976). In this case, the likelihood simplifies to

$$\begin{aligned} p(T_{obs} = t_{obs}, W = w|\eta, \theta) &= \sum_{t_{miss}} p(W = w|T_{obs} = t_{obs}, \theta) \frac{1}{c(\eta, \mathcal{T})} e^{\eta \cdot g((t_{obs}, t_{miss}))} \\ &= p(W = w|T_{obs} = t_{obs}, \theta) \sum_{t_{miss}} \frac{1}{c(\eta, \mathcal{T})} e^{\eta \cdot g((t_{obs}, t_{miss}))} \\ &\propto \sum_{t_{miss}} \frac{1}{c(\eta, \mathcal{T})} e^{\eta \cdot g((t_{obs}, t_{miss}))}. \end{aligned} \quad (3)$$

Thus, when the sampling process is ignorable, inferences on η are not affected by $p(W = w|T_{obs} = t_{obs}, \theta)$, and so knowledge of the sampling process is not essential for the process of inference.

Having defined the full and observed likelihood, it is also useful to define the *missing data likelihood*:

$$p(T_{miss} = t_{miss}|W = w, T_{obs} = t_{obs}, \eta, \theta) = \frac{p(W = w|T = (t_{obs}, t_{miss}), \theta) e^{\eta \cdot g((t_{obs}, t_{miss}))}}{c(t_{obs}, w, \eta, \theta)}$$

where

$$c(t_{obs}, w, \eta, \theta) = \sum_{t_{miss}} p(W = w|T = (t_{obs}, t_{miss}), \theta) e^{\eta \cdot g((t_{obs}, t_{miss}))}.$$

The (observed data) likelihood can then be rewritten as the ratio of two normalising constants

$$\begin{aligned} p(T_{obs} = t_{obs}, W = w|\eta, \theta) &= \frac{1}{c(\eta, \mathcal{T})} \sum_{t_{miss}} p(W = w|T = (t_{obs}, t_{miss}), \theta) e^{\eta \cdot g((t_{obs}, t_{miss}))} \\ &= \frac{c(t_{obs}, w, \eta, \theta)}{c(\eta, \mathcal{T})}, \end{aligned}$$

and using this, we may write the observed data log likelihood ratio of (η, θ) versus (η_0, θ_0)

as

$$\begin{aligned}
 \ell(\eta, \theta) - \ell(\eta_0, \theta_0) &= \log\left(\frac{c(t_{obs}, w, \eta, \theta)}{c(t_{obs}, w, \eta_0, \theta_0)}\right) - \log\left(\frac{c(\eta, \mathcal{T})}{c(\eta_0, \mathcal{T})}\right) \\
 &= \log\left(\sum_{t_{miss}} \frac{p(W = w|T = t, \theta)}{p(W = w|T = t, \theta_0)} e^{(\eta - \eta_0) \cdot g(t)} \frac{p(W = w|T = t, \theta_0) e^{\eta_0 \cdot g(t)}}{c(t_{obs}, w, \eta_0, \theta_0)}\right) \\
 &\quad - \log\left(\sum_{t_{miss}} e^{(\eta - \eta_0) \cdot g(t)} \frac{e^{\eta_0 \cdot g(t)}}{c(\eta, \mathcal{T})}\right) \\
 &= \log(E_{\eta_0, \theta_0}\left(\frac{p(W = w|T, \theta)}{p(W = w|T, \theta_0)} e^{(\eta - \eta_0) \cdot g(T)}\right) | W = w, T_{obs} = t_{obs}) \quad (4) \\
 &\quad - \log(E_{\eta_0}(e^{(\eta - \eta_0) \cdot g(T)})) \\
 &= \log(E_{\eta_0, \theta_0}(e^{(\eta - \eta_0) \cdot g(T)} | T_{obs} = t_{obs})) - \log(E_{\eta_0}(e^{(\eta - \eta_0) \cdot g(T)})) \\
 &\quad + \log\left(\frac{E_{\eta, \theta}(P(W = w|T, \theta) | T_{obs} = t_{obs})}{E_{\eta_0, \theta_0}(P(W = w|T, \theta_0) | T_{obs} = t_{obs})}\right). \quad (5)
 \end{aligned}$$

Both equation (4) and equation (5) motivate algorithms to draw inferences about η and θ . Section 4 describes the algorithm motivated by equation (4), and Appendix A.1 outlines an algorithm using equation (5).

4. Calculating the MLE with MCMC

For most models, equation (4) is not analytically solvable. However we may approximate it by Markov Chain Monte Carlo (MCMC). Let $t^{(i)}$ and $t_m^{(i)}$ where $i \in (1, \dots, M)$ be samples from the full likelihood and missing data likelihood respectively with parameters η_0, θ_0 . Then equation (4) may be approximated by

$$\ell(\eta, \theta) - \ell(\eta_0, \theta_0) \approx \log\left(\frac{1}{M} \sum_i \frac{p(w|t_m^{(i)}, \theta)}{p(w|t_m^{(i)}, \theta_0)} e^{(\eta - \eta_0) \cdot g(t_m^{(i)})}\right) - \log\left(\frac{1}{M} \sum_i e^{(\eta - \eta_0) \cdot g(t^{(i)})}\right) \quad (6)$$

As η, θ move away from η_0, θ_0 the quality of this approximation degrades. Because we will be optimising equation (4), it is useful to have both the first and second derivatives of the log likelihood, which are

$$\begin{aligned}
 \frac{\delta \ell}{\delta \eta} &= E_{\eta, \theta}(g_i(t) | T_{obs} = t_{obs}, W = w) - E_{\eta, \theta}(g_i(T)) \\
 \frac{\delta^2 \ell}{\delta \eta_i \delta \eta_j} &= -\text{cov}(g_i(T), g_j(T)) + \text{cov}(g_i(T), g_j(T) | T_{obs} = t_{obs}, W = w).
 \end{aligned}$$

The expectations and covariances in these derivatives can be approximated using the conditional and unconditional MCMC samples and thus we can then use the following algorithm to approximate the MLE.

- (a) Let $k = 0$ and choose initial parameter values $\eta^{(0)}, \theta_0$.
- (b) Use MCMC to generate k samples, $t_{miss}^{(i)}$ from $P(T_{miss} = t_{miss} | \eta^k, T_{obs} = t_{obs}, W = w)$.

- (c) Use MCMC to generate m samples $t^{(i)}$ from $P(T = t|\eta^k)$.
- (d) Using the samples from step 2 and 3 in equation (6), find η^{k+1}, θ^{k+1} maximising the likelihood ratio, subject to $\|\eta^{k+1} - \eta^k\| < \epsilon$ and $\|\theta^{k+1} - \theta^k\| < \epsilon$.
- (e) If the likelihood has not converged, set $k = k + 1$ and go to step 2.
- (f) Let the MLE estimate be $\hat{\eta} = \eta^{k+1}$ and $\hat{\theta} = \theta^{k+1}$

Asymptotic standard errors for $\hat{\eta}$ may be obtained using an MCMC approximation to the Fisher information (i.e. the second derivative of the log likelihood). While asymptotics of the Fisher information are not assured with respect to ERNM (or ERGM) models, Fellows and Handcock (2012) show strong empirical agreement between the Fisher information standard errors and parametric bootstrap simulations. Standard errors for the mean value parameters $\hat{\mu} = E(g(T)|\eta = \hat{\eta})$ can be approximated by MCMC sampling.

5. Specific forms of partial observation

In this section we consider the three common forms of partial observation considered in the introduction, each corresponding to a different mechanism of partial observation or conceptualisation of that mechanism.

5.1. Missing Data: Unobserved Relational Information

It is common when surveying networked populations that there are insufficient resources to conduct a census of the population and their relations. For efficiency reasons, a sampling based survey is undertaken, or the full network is partially observed due to non-response. In this sub-section, we give an illustration of the effect of non-response where the dyad information is missing completely at random. We consider the relations of “liking” among 18 monks in a monastery (Sampson, 1969). The network analysed has a directed edge between two monks if the sender monk ranked the receiver monk in the top three monks for positive affection in any of the three interviews given over a twelve month period (Hoff et al., 2002). The sociogram of this data-set is shown in Figure 1. One nodal attribute of interest is an indicator of attendance at the minor “Cloisterville” seminary before coming to the monastery.

We fit a simple homophily model on Cloisterville status using the full data. We then ran simulations on the effect of missingness by selecting dyads, and Cloisterville status variates, completely at random and setting them to missing. Figure 2 shows one simulated missingness pattern with 15% missing. We ran 100 simulations at each missingness percentage. Means and standard deviations of the ERNM models fit to these simulated missingness patterns are displayed in Figure 3.

We see that the standard deviations of the estimates increase as the amount of missingness increases. At the higher missingness levels some bias is apparent relative to the full data MLE, but not more than one standard deviation. One possible explanation for this bias is that there were only six monks who attended Cloisterville, and so at 50% missingness, a significant number of samples will include no (or perhaps a single) Cloisterville monks.

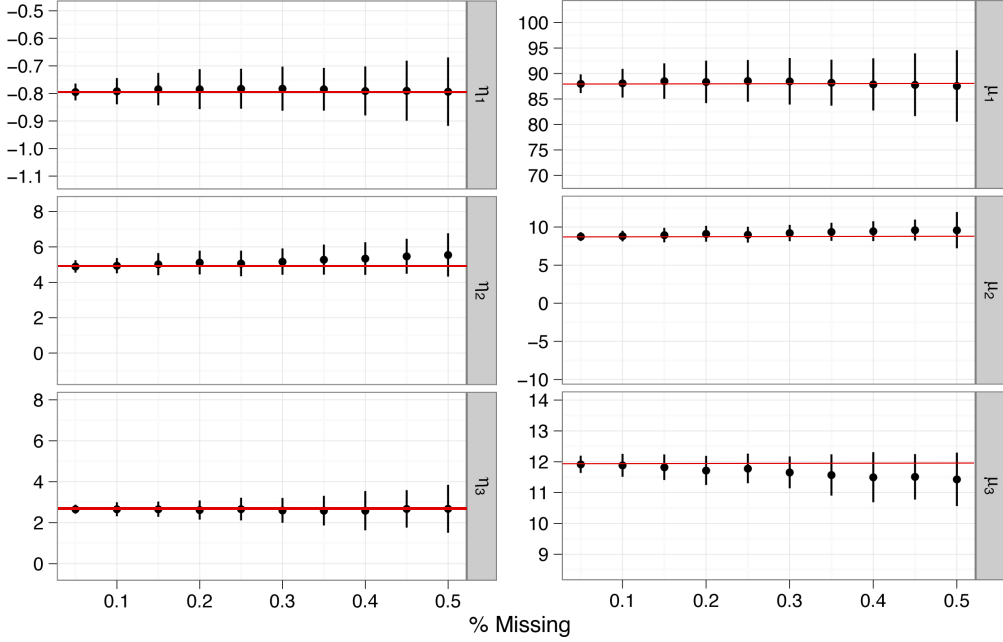


Fig. 3. Means and standard deviations of model estimates. Red lines indicate fully observed MLE

5.2. Latent Variables: Stochastic Block Models

In this sub-section we consider the situation where some characteristics of the network are posited but unobserved. Specifically, we consider the case where each node of the network belongs to a latent class, and the structure of the network depends on that latent class. The traditional approach to this has been stochastic block models Nowicki and Snijders (2001), and here we show how these models fall naturally out of our general formulation.

It is apparent from Figure 1 that the pattern of “liking” between the monks may exhibit clustering. Through close sociological study, Sampson (1969) identified three clusters which he dubbed the Turks, Loyal Opposition and the Outcasts (see: Figure 1). Here we will attempt to identify clusters by inferring class membership from the graph. We fit the simple homophily model of Section 2.1 to this data, assuming a class covariate, X , with three levels, and that all of the monks are “missing” their class covariate. The simple homophily model treated this way represents a novel latent block model in the spirit of Nowicki and Snijders (2001). Note that the missingness process here is ignorable because it does not depend on unobserved quantities as all of the x values are missing regardless of the Y values. We fit the model using the algorithm in Section 4. Table 1 shows the maximum likelihood parameter estimates, along with standard errors of the estimators based on the Fisher information.

The natural parameter estimates indicate significant homophily in tie formation based on the class. It also indicates that the number of monks in the third class is significantly more than those of the other two classes, which are not statistically significantly different in size. The mean value parameters indicate that the expected number of ties is about 88,

Table 1. Latent Class model for Sampson’s monks.

Term	$\hat{\eta}$	$\hat{\mu}$	s.e.($\hat{\eta}$)	s.e.($\hat{\mu}$)
# of edges	-0.58	88.23	0.14	7.48
Homophily	7.28	15.30	0.91	1.33
# in group 0	-2.50	3.95	1.44	1.08
# in group 1	-0.02	6.95	1.31	0.99

and the expected numbers in the three groups are 4, 7 and 7.

An advantage of this approach is that we can investigate the probability of class membership, which is well defined through our framework as $p(X = x|Y = y_{obs}, \eta)$. To compute $p(X = x|Y = y_{obs}, \eta)$ we simulated a large number of samples from $p(X = x|Y = y_{obs}, \hat{\eta})$ using MCMC to show the probability of the monks being in the classes displayed in Figure 1 to be above 0.9999. These clusters were also identical to those chosen by Sampson (1969) and verified by later research Breiger et al. (1975); Handcock et al. (2006).

In addition to assuming a set number of latent classes for the model, we can also use the MLE procedure to select an appropriate number of clusters for the data. We fit the simple homophily model with a latent variable X able to take a potentially large number of values (e.g., the number of monks). In this case $p(X = x|Y = y_{obs}, \hat{\eta})$ places zero mass for all but three of the groups. This is evidence that the three groups we have identified are a good classification for these data. More sophisticated model selection approaches for choosing the number of clusters are possible (Handcock et al., 2006), and are left for future work.

Our form of the stochastic block model is conceptually very clean with the ability to naturally incorporate additional covariates, multiple membership variables, and extensions to an unbounded numbers of classes. Inference is straightforward, and quantities such as the probability of class membership are well defined and interpretable. We leave a full exploration of these for latter work.

5.3. Network Sampling: Biased Seed Link-Tracing

Handcock and Gile (2010) explored the idea of sampling networks by tracing the edges. As a general concept, link tracing involves selecting one or more seed nodes, and then observing the edges connected to those seeds. One or more of these edges are then followed to the neighbouring node, whose ties are observed, and the process is continued. Each iteration of this process is known as a wave.

Provided that the seed nodes are chosen at random, and the method by which edges are chosen to be followed depends only on the observed data, this missingness process is ignorable. To be explicit, consider a link tracing process with k waves. Let w_i be the ordered set of nodes and edges sampled in the i th wave in the order in which they were sampled, $w = \{w_0, \dots, w_k\}$, and $w_{-i} = \{w_0, \dots, w_{i-1}, w_{i+1}, \dots, w_k\}$. If the seeds are chosen at random, and the edges followed by the sampling process are also chosen at random, then $p(W = w|T = t, \theta) = p(W = w|T_{obs} = t_{obs}, \theta)$, implying that the missingness is ignorable.

In many cases, however, the seeds are not chosen at random from the population, but are some form of convenience sample. For example, in a population where some people have an infection and others do not, we may start with a sample of s_i seeds picked at random from among the infected individuals, and s_{-i} seeds picked from the non-infected individuals. These seeds are then used as a starting point for standard link tracing. We may then write

the sampling probability as

$$\begin{aligned} p(w|t, \theta) &= p(w_0|t, \theta)p(w_{-0}|t_{obs}, w_0, \theta) \\ &= \frac{(n_i - s_i)!}{n_i!} \frac{(n_{-i} - s_{-i})!}{n_{-i}!} p(w_{-0}|t_{obs}, w_0, \theta), \end{aligned}$$

where n_i and n_{-i} are the number of infected and non-infected in the population, respectively. Note that $p(w_{-0}|t_{obs}, w_0, \theta)$ does not depend on t_{miss} and may be factored out of the likelihood in equation (2). Thus there is no need to calculate $p(w_{-0}|t_{obs}, w_0, \theta)$ explicitly, as it makes no impact on the likelihood. Hence, in this case, we can compute the likelihood without knowing the specific mechanism of seed selection.

5.4. Network Sampling: Positive Contact Tracing

As emerging epidemics develop, control measures (e.g., treatment, isolation and culling) focus on those members of the population that are known to have the infection. Because there are often many infected people who are unobserved, control can be ineffective (e.g., HIV (Potterat et al., 1989)). The alternative of applying control measures to the entire population can be economically infeasible or ineffective (e.g., some instances of safe sex education) (Potterat et al., 1989; Klinkenberg et al., 2006). Contact tracing is the hybrid approach of treating both the known infected individuals and those who may have been infected by them (Potterat et al., 1989; Klinkenberg et al., 2006). In U.S. public health, health clinics are required by state law to notify those at risk from infection due to their sexual relations with individuals tested, and found to be infected, by the clinic. The process of locating, notifying and then testing partners that may have been exposed to an infectious agent allows additional information about the partners to be collected. While the primary purpose of contact tracing is disease control via partner notification and partner services, it is also a form of data collection that is rarely utilised. Such approaches are used most commonly for syphilis and HIV/AIDS, but also for other STIs such as gonorrhea and chlamydia (Golden et al., 2004), as well as routinely for tuberculosis and infectious disease outbreaks. Contact tracing has also been applied in many recent epidemics (Fenner et al., 1988; Ferguson et al., 2001; Donnelly et al., 2003). In *positive contact tracing*, we follow all edges from infected nodes, but edges from uninfected nodes are not followed.

While the process varies from state to state and also by disease, we consider the following biased seed link tracing process:

- (a) Select s_{-i} seed subjects at random from among the non-infected population, observe them.
- (b) Select s_i seeds subjects at random from among the infected population, observe them.
- (c) Choose the next infected seed at random.
- (d) Observe all edges from the selected subject, and the infection status of these subjects.
- (e) For all infected neighbours of the selected subject, go to step 4.
- (f) If all the seeds have not been chain sampled, go to step 3

We simulated a networked population of $n = 1000$ people from the simple homophily model of Section 2.1 with natural parameters of $\eta = (-5.8, .7, -1.95)$. The number of infected nodes was fixed at 150. The generated network had a mean degree of 3.1, and its degree distribution is displayed in Figure 4. There were 296 infected to non-infected ties, with the mixing distribution displayed in Figure 5 indicating moderate homophily.

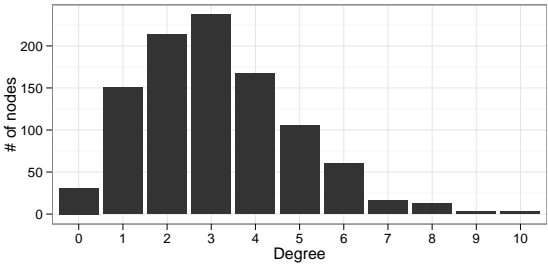


Fig. 4. Degree distribution of the networked population.

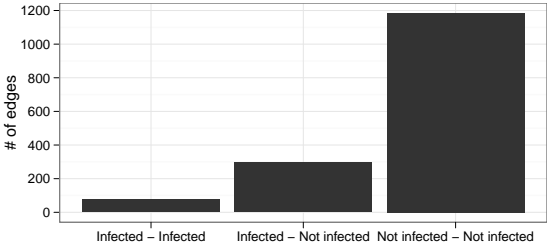


Fig. 5. Mixing statistics: Counts of the numbers of edges by the infection status of the incident nodes for the networked population.

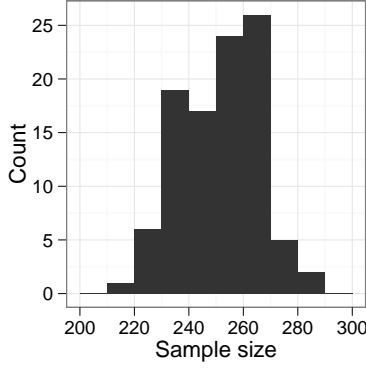


Fig. 6. Sizes of the contact-traced samples based on 40 seed subjects ($s_i = 40, s_{-i} = 0$).

Starting with $s_i = 40$ infected seeds, we simulated 100 positive link tracing samples for each of $s_{-i} = (0, 45, 90, 135, 180, 225)$. Figure 6 displays a histogram of the sizes of the samples when there are no non-infected seeds (i.e., $s_{-i} = 0$).

To provide a comparison for our method we considered two estimators that could be utilised. Neither of them uses a model for the networked population but is motivated by approximations to the sampling design. The first treats the sample as a simple random sample

$$\text{Naive} = n \frac{n_i}{n_i + n_u},$$

where n_i and n_u are the number of infected and uninfected in the sample respectively. The second adjusts for the sampling of the seeds

$$\text{Naive (seed adj.)} = (n - s_i - s_{-i}) \frac{n_i - s_i}{n_i - s_i + n_u - s_{-i}} + s_i.$$

Our approach is to fit an ERNM to the contact tracing data. In this situation the contact tracing sampling design is clearly informative. For comparison, we compute two estimates of the model. The first takes into account the informativeness of the contact tracing design (MNAR) and the other assume it is ignorable (MAR). These are based on the likelihoods 2 and 5, respectively, and the algorithm in Section 4.

Figure 7 shows the results for each of the estimators over the samples. The median of the MNAR estimator is centred around the true value of 150 in all sampling scenarios, while the MAR estimator performs poorly with all infected seeds ($s_{-i} = 0$) and increasingly well as the number of non-infected seeds increases to $s_{-i} = 225$. This is somewhat expected as the proportion of infected in the seeds approximately matches that of the population when $s_{-i} = 225$. The two naive estimators are significantly biased across all samples. This is especially true for the sample mean which is biased both by the seed selection and by the link-tracing design. The adjusted sample mean corrects somewhat for the seed bias but does not represent the link-tracing.

This application illustrates the advantage of the model-based approach over the *ad hoc* estimators. By representing the structure of the networked population, the model-based

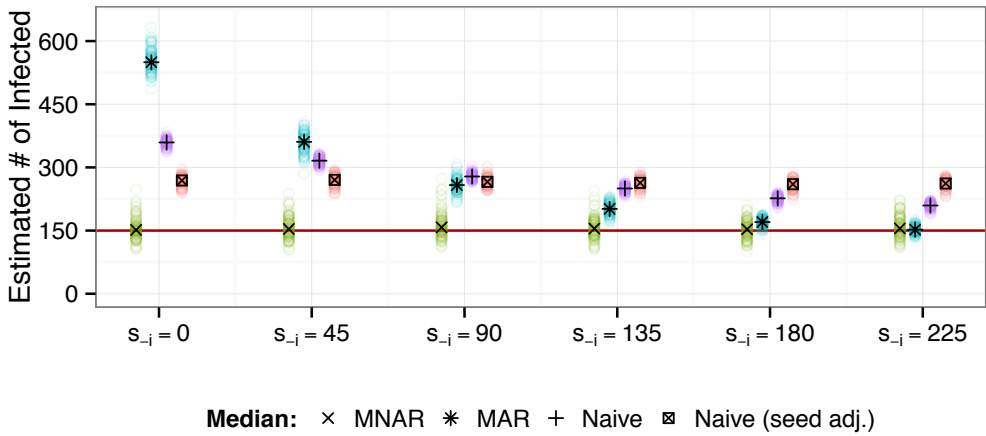


Fig. 7. Estimates via contact tracing with $s_i = 40$ infected seeds and varying numbers of non-infected seeds.

approach can leverage the information in the data more efficiently.

6. Discussion

In this paper we have given a concise and systematic statistical framework for dealing with partially observed network data when some knowledge is available on the sampling design. The framework includes, but is not restricted to, ignorable sampling designs. We have also shown that likelihood-based inference is practical under partial observation for ERN models, and that the likelihood framework naturally accommodates standard sampling designs.

We developed and implemented algorithms to compute Monte Carlo approximations to the likelihood, and showed how these can be used in practice. Three important special cases of these designs were demonstrated in Section 5. In Sub-section 5.1 we consider a missingness process which randomly selected dyads and nodal attributes to be missing. Sub-section 5.1 considers the case where *all* nodal attributes are missing, thus introducing a novel form of the latent cluster model.

In Sub-section 5.3 we consider non-ignorable sampling in the context of contact tracing data, a case of vital importance to public health. At present, this is the first statistically defensible approach to inference in this form of data. The example presented here shows that the MLE estimation task is robust, in that it can be applied successfully to moderately large networks (1000 nodes), with significant missingness (70% of nodes unobserved), but is limited by the fact that inference was performed on a simulated network. Whether the model presented here would provide a good fit for real public health data remains an important research question that we hope to address in the future.

Appendix: Algorithmic and Computational Details

A.1: Alternate MLE Formulation

While the algorithm outlined in Section 4 works well, there are some situations where an alternate formulation using equation (5) may be useful. First let us consider the case where $\theta = \theta_0$, then the likelihood is

$$\ell(\eta) - \ell(\eta_0) = \log(E_{\eta_0}(e^{(\eta - \eta_0) \cdot g(T)} | t_{obs})) - \log(E_{\eta_0}(e^{(\eta - \eta_0) \cdot g(T)})) + \log\left(\frac{E_{\eta}(P(W = w | T, \theta) | T_{obs} = t_{obs})}{E_{\eta_0}(P(W = w | T, \theta) | T_{obs} = t_{obs})}\right) \quad (7)$$

The first expectation, and the expectation in the denominator of the third term, can be calculated using an MCMC sample from $p(t | t_{obs}, \eta_0)$. The second can be approximated with an MCMC sample from $p(t | \eta_0)$. The numerator of the third term can be approximated by importance sampling.

$$E_{\eta}(P(W = w | T, \theta) | T_{obs} = t_{obs}) \approx \frac{1}{k} \sum_i^k p(w | t^{(i)}, \theta) \omega^{(i)}$$

where $t^{(i)} \sim p(t | t_{obs}, \eta_0)$ and

$$\omega^{(i)} = \frac{e^{(\eta - \eta_0) \cdot g(t^{(i)})}}{\sum_j^k e^{(\eta - \eta_0) \cdot g(t^{(j)})}}$$

If the sampling process is ignorable, then the third term drops out of the likelihood ratio. The first and second derivatives of the likelihood are useful in the maximisation process. For notational convenience let $\Delta_i(t) = g_i(t) - E(g_i(T))$.

$$\begin{aligned} \frac{\delta \ell}{\delta \eta} &= \frac{\delta}{\delta \eta_i} \log\left(\sum_{t_{miss}} p(W = w | T = t) P(T_{miss} = t_{miss} | \eta, T_{obs} = t_{obs}) P(T_{obs} = t_{obs} | \eta)\right) \\ &= \frac{\sum_{t_{miss}} p(W = w | T = t) \Delta_i(t) P(T_{miss} = t_{miss} | \eta, T_{obs} = t_{obs}) P(T_{obs} = t_{obs} | \eta)}{\sum_{t_{miss}} p(W = w | T = t) P(T_{miss} = t_{miss} | \eta, T_{obs} = t_{obs}) P(T_{obs} = t_{obs} | \eta)} \\ &= \frac{E(p(W = w | T) \Delta_i(T) | T_{obs} = t_{obs})}{E(p(W = w | T) | T_{obs} = t_{obs})} \end{aligned}$$

$$\begin{aligned} \frac{\delta^2 \ell}{\delta \eta_i \delta \eta_j} &= \frac{\delta}{\delta \eta_j} \frac{\sum_{t_{miss}} P(W = w | T = t) \Delta_i(t) P(T_{miss} = t_{miss} | \eta, T_{obs} = t_{obs}) P(T_{obs} = t_{obs} | \eta)}{\sum_{t_{miss}} P(W = w | T = t) P(T_{miss} = t_{miss} | \eta, T_{obs} = t_{obs}) P(T_{obs} = t_{obs} | \eta)} \\ &= -\text{cov}(g_i(T), h_j(T)) + \frac{E(p(W = w | T) \Delta_i(T) \Delta_j(T) | T_{obs} = t_{obs})}{E(p(W = w | T) | T_{obs} = t_{obs})} \\ &\quad - \frac{E(p(W = w | T) \Delta_i(T) | T_{obs} = t_{obs}) E(p(W = w | T) \Delta_j(T) | T_{obs} = t_{obs})}{E(p(W = w | T) | T_{obs} = t_{obs})^2} \end{aligned}$$

And if the missingness process is ignorable, these equations simplify to

$$\frac{\delta \ell}{\delta \eta} = E(\Delta_i(T) | T_{obs} = t_{obs})$$

$$\frac{\delta^2 \ell}{\delta \eta_i \delta \eta_j} = -\text{cov}(g_i(T), g_j(T)) + \text{cov}(g_i(T), g_j(T) | T_{obs} = t_{obs})$$

If we fix η , then the observed likelihood of θ

$$\begin{aligned} L(\theta | t_{obs}, w, \eta) &\propto P(t_{obs} | \eta) E(P(W = w | T, \theta) | T_{obs} = t_{obs}) \\ &= E(P(W = w | T, \theta) | T_{obs} = t_{obs}, \eta) \end{aligned}$$

can be maximised to find the MLE of θ .

This motivates the following algorithm for maximising the observed data likelihood.

- (a) Let $k = 0$ and choose initial parameter values $\eta^{(0)}$, θ_0 .
- (b) Use MCMC to generate k samples, $t_{miss}^{(i)}$ from $P(t_{miss} | \eta^k, t_{obs})$.
- (c) Use MCMC to generate m samples $t^{(i)}$ from $P(t | \eta^k)$.
- (d) Set $\theta^{k+1} = \text{argmax}(E(P(w | T, \theta) | T_{obs} = t_{obs}, \eta))$, with samples from step 2 used to approximate the expectation.
- (e) Using the samples from steps 2 and 3 to approximate the relevant expectations, find η^{k+1} maximising equation (7) subject to $\|\eta^{k+1} - \eta^k\| < \epsilon$.
- (f) Set $k = k + 1$, and go to step 2.

The disadvantage of this method is that if the networks generated by the MNAR process are very different from those generated assuming MAR, the estimates of the last expectation in equation (7) can become unstable. The benefit of using this method is that the sampling probability ($P(W = w | T = t, \theta)$) only needs to be calculated for networks included in the sample, and not at every MCMC step as is required by the algorithm in Section 4, so if the sampling probability is computationally expensive to calculate, this method can be significantly faster than the one outlined in Section 4

A.2: Estimating Network Statistics

We can use MCMC samples from $p(t_{miss} | t_{obs}, \eta)$ to estimate the network statistics of the sampled network. Suppose that we have used MCMC to draw k samples $t_{miss}^{(i)}$ from the distribution $p(t_{miss} | t_{obs}, \eta)$, and $t^{(i)} = (t_{obs}, t_{miss}^{(i)})$. Then we can estimate the expectation of a set of network statistics g as

$$E(g(T) | t_{obs}, \eta) \approx \frac{1}{k} \sum_{i=0}^k g(t^{(i)}).$$

However, this equation ignores the possible bias introduced by our sampling process w . The distribution that we should be sampling from is the full conditional distribution of t_{miss} ,

$$p(T_{miss} = t_{miss} | T_{ob} = t_{obs}, W = w, \eta) \propto p(T_{miss} = t_{miss} | T_{obs} = t_{obs}, \eta) p(W = w | T = t, \theta).$$

We then use importance sampling to estimate the relevant quantity

$$E(g(T) | t_{obs}, w, \eta, \theta) \approx \frac{\sum_{i=0}^k g(t^{(i)}) p(W = w | T = t^{(i)}, \theta)}{\sum_{i=0}^k p(W = w | T = t^{(i)}, \theta)}.$$

References

- Breiger, R. L., Boorman, S. A., and Arabie, P. (1975). An algorithm for clustering relational data, with applications to social network analysis and comparison with multidimensional scaling. *Journal of Mathematical Psychology*, 12, 328–383.
- Donnelly, C. A., Ghani, A. C., Leung, G. M., et al, and Anderson, R. M. (2003). Epidemiological determinants of spread of causal agent of severe acute respiratory syndrome in Hong Kong. *Lancet*, 361(9371), 1761–1766.
- Fellows, I., and Handcock, M. S. (2012). Exponential-family Random Network Models. *ArXiv e-prints*.
- Fenner, F., Henderson, D. A., Arita, I., Jezek, Z., and Ladnyi, I. (1988). Smallpox and its eradication. Tech. rep., Geneva: World Health Organization.
- Ferguson, N. M., Donnelly, C. A., and Anderson, R. M. (2001). Transmission intensity and impact of control policies on the foot and mouth epidemic in Great Britain. *Nature*, 413(6855), 542–548.
- Frank, O., and Strauss, D. (1986). Markov graphs. *Journal of the American Statistical Association*, 81(395), 832–842.
- Gile, K. J. (2011). Improved inference for respondent-driven sampling data with application to HIV prevalence estimation. *Journal of the American Statistical Association*, 106(493), 135–146.
- Gile, K. J., and Handcock, M. S. (2010). Respondent-driven sampling: An assessment of current methodology. *Sociological Methodology*, 40, 285–327.
URL <http://arxiv.org/abs/0904.1855v1>
- Gile, K. J., and Handcock, M. S. (2011). Network model-assisted inference from respondent-driven sampling data. *ArXiv Preprint*.
URL <http://arxiv.org/abs/1108.0298>
- Golden, M. R., Hogben, M., Potterat, J. J., and Handsfield, H. H. (2004). HIV partner notification in the United States: a national survey of program coverage and outcomes. *Sex Transm Dis*, 31(12), 709–712.
- Handcock, M. S., and Gile, K. J. (2010). Modeling networks from sampled data. *Annals of Applied Statistics*, 272(2), 383–426.
- Handcock, M. S., Raftery, A. E., and Tantrum, J. M. (2006). Model-based clustering for social networks. *Journal of the Royal Statistical Society Series A*, 170, 1–22.
- Harris, K. M., Florey, F., Tabor, J., Bearman, P. S., Jones, J., and Udry, J. R. (2003). The national longitudinal of adolescent health: Research design [WWW document]. Tech. rep., Carolina Population Center, University of North Carolina at Chapel Hill, Available at: <http://www.cpc.unc.edu/projects/addhealth/design>.
- Hoff, P. D., Raftery, A. E., and Handcock, M. S. (2002). Latent space approaches to social network analysis. *Journal of the American Statistical Association*, 97(460), 1090–1098.

- Hunter, D. R., and Handcock, M. S. (2006). Inference in curved exponential family models for networks. *Journal of Computational and Graphical Statistics*.
- Klinkenberg, D., Fraser, C., and Heesterbeek, H. (2006). The effectiveness of contact tracing in emerging epidemics. *PLoS ONE*, 1(1), e12.
- Nowicki, K., and Snijders, T. A. B. (2001). Estimation and prediction for stochastic block-structures. *Journal of the American Statistical Association*, 96(455), 1077–1087.
- Potterat, J. J., Spencer, N. E., Woodhouse, D. E., and Muth, J. B. (1989). Partner notification in the control of human immunodeficiency virus infection. *American Journal of Public Health*, 79(7), 874–876.
- Rubin, D. (1976). Inference and missing data. *Biometrika*, 63, 581–592.
- Sampson, S. F. (1969). *Crisis in a Cloister*. PhD in Sociology, Cornell University.
- Udry, J. R. (2003). The national longitudinal of adolescent health: (add health), waves I and II, 1994-1996; wave III, 2001-2002 [machine-readable data file and documentation]. Tech. rep., Carolina Population Center, University of North Carolina at Chapel Hill.